



## Correspondence

# Predictors of granulomatous lymphocytic interstitial lung disease in common variable immunodeficiency: Methodological issues



We read with great interest the article by Hartono et al<sup>1</sup> that was published in the *Annals of Allergy, Asthma & Immunology* in May 2017. They intended to determine a set of clinical and/or laboratory parameters associated with granulomatous lymphocytic interstitial lung disease (GLILD). Based on univariable analysis, they stated that splenomegaly was independently associated with GLILD (odds ratio [OR] 17.3, 95% confidence interval [CI] 3.9–74.5). Although the study was useful and very interesting, some methodologic issues need to be considered.

First, Greenland et al believe that the large effect size estimate such as the OR with considerably wide CIs not only does not indicate strong association between exposure and outcome variables but also represents a large error.<sup>2,3</sup> In fact, they stated that large effect size estimates with wide CIs resulted in sparse data in which the number of observations on the association between exposure and outcome variables was not sufficient.<sup>3</sup> The effect size estimate would be inflated and its CI would widen because of the sparse data bias.<sup>3–5</sup> Fortunately, this bias is preventable and correctable using modern advanced methods such as penalization via data augmentation.<sup>2,3</sup> We explored the data provided in the study by Hartono et al and found that the number of observations on the association between splenomegaly and GLILD was insufficient ( $n_{\text{Exp} + \text{cases}} = 18$ ;  $n_{\text{Exp} - \text{cases}} = 3$ ;  $n_{\text{Exp} + \text{controls}} = 8$ ;  $n_{\text{Exp} - \text{controls}} = 23$ ) and that sparse data bias was unavoidable.<sup>1</sup> Hence, we re-estimated the OR using the penalization method with log-F prior distribution (3.9, 3.9) and found that the OR decreased and its CI narrowed (OR 8.20, 95% CI 2.48–27.12).

Second, Hartono et al indicated that splenomegaly was independently associated with GLILD, which is problematic.<sup>6</sup> They did not conduct multivariable models to control potential confounding variables. Hence, it is not clear whether the splenomegaly is independently associated with GLILD or whether the association might be confounded by potential confounding variables.

The take home message for readers is that appropriate advanced statistical methods should be applied to sparse data to remove or decrease the sparse data bias as much as possible. Also, multivariable models should be applied to examine whether the exposure of interest is independently associated with the outcome.

Erfan Ayubi, PhD\*

Saeid Safiri, PhD<sup>†,‡</sup>

\*Department of Epidemiology  
School of Public Health  
Shahid Beheshti University of Medical Sciences  
Tehran, Iran

†Managerial Epidemiology Research Center  
Department of Public Health  
School of Nursing and Midwifery  
Maragheh University of Medical Sciences  
Maragheh, Iran

‡Department of Epidemiology and Biostatistics  
School of Public Health  
Tehran University of Medical Sciences  
Tehran, Iran  
saeidsafiri@gmail.com

## References

- [1] Hartono S, Motosue MS, Khan S, et al. Predictors of granulomatous lymphocytic interstitial lung disease in common variable immunodeficiency. *Ann Allergy Asthma Immunol*. 2017;118:614–620.
- [2] Greenland S, Mansournia MA. Penalization, bias reduction, and default priors in logistic and related categorical and survival regressions. *Stat Med*. 2015;34:3133–3143.
- [3] Greenland S, Mansournia MA, Altman DG. Sparse data bias: a problem hiding in plain sight. *BMJ*. 2016;352:i1981.
- [4] Ayubi E, Safiri S. Bias in association between FEV1/FVC% predicted at 7 years and asthma-COPD overlap syndrome [published online ahead of print March 31, 2017]. *Am J Respir Crit Care Med*. <http://dx.doi.org/10.1164/rccm.201702-0370LE>.
- [5] Ayubi E, Safiri S. Lateral lymph node recurrence after total thyroidectomy and central neck dissection in patients with papillary thyroid cancer without clinical evidence of lateral neck metastasis: Comment on data sparsity. *Oral Oncol*. 2017;69:128.
- [6] Rothman KJ, Greenland S, Lash TL. *Modern Epidemiology*. Baltimore, MD: Lippincott Williams & Wilkins; 2008.

**Disclosures:** Authors have nothing to disclose.

## Author's response

We would like to thank authors for their comments on our article,<sup>1</sup> which are both valid and valuable to consider. We agree that the estimated odds ratios are biased away from the null hypothesis given the sparse data. We also agree that the use of the word

*independently* is incorrect in the abstract. The odds ratios reported in the abstract and in table 4 are univariate odds ratios. The term *independently associated* should be reserved for results from a multivariable analysis in which we have adjusted for the other



variables in the model. In a case-control study, it is appropriate to estimate associations in a multivariable model, but it is not appropriate to actually present the mathematical formula for a prediction model without taking into account the true prevalence of the disease. Because we chose to study an equal number of controls as cases, the prevalence of the disease is preset at 50%, which affects the intercept estimate in our multivariable logistic model. With just 26 cases and 21 controls, the multivariable logistic model that was fit to derive the receiver operating characteristic curve was considerably overfit.<sup>2</sup>

Finally, the commentary authors are suggesting a Bayesian approach<sup>3</sup> to penalize or shrink the estimates of the odds ratios. We, as nonstatisticians, did not apply Bayesian methods and used a more simplistic approach by reporting the univariate results. We acknowledge that the estimated odds ratios are biased away from the null hypothesis given the sparse data. However, to apply such a penalization the research team needs to have some background knowledge of what to anticipate for the 95% confidence interval for the effect size. The authors of the commentary assume that the prior 95% confidence intervals for the odds ratio would range from 0.1 to 10 and thereby weighted the data accordingly to perform the penalization. We plan to undertake a similar study using the United States Immunodeficiency Network data set<sup>4</sup> to obtain more accurate estimates by incorporating background information to penalize or shrink the estimated odds ratios.

Stella Hartono, MD, PhD\*  
Megan S. Motosue, MD†

Shakila Khan, MD‡  
Vilmarie Rodriguez, MD‡  
Vivek N. Iyer, MD§  
Rohit Divekar, MBBS, PhD†  
Avni Y. Joshi, MD, MSc†,‡

\*Mayo Medical School  
Rochester, Minnesota

†Division of Allergic Diseases  
Mayo Clinic Children's Center  
Rochester, Minnesota

‡Department of Pediatric and Adolescent Medicine  
Mayo Clinic College of Medicine  
Rochester, Minnesota

§Division of Pulmonary and Critical Care Medicine  
Mayo Clinic  
Rochester, Minnesota  
[joshi.avni@mayo.edu](mailto:joshi.avni@mayo.edu)

## References

- [1] Hartono S, Motosue MS, Khan S, et al. Predictors of granulomatous lymphocytic interstitial lung disease in common variable immunodeficiency. *Ann Allergy Asthma Immunol*. 2017;118:614–620.
- [2] Zhang Z. Too much covariates in a multivariable model may cause the problem of overfitting. *J Thorac Dis*. 2014;6:E196–E197.
- [3] Greenland S, Mansournia MA. Penalization, bias reduction, and default priors in logistic and related categorical and survival regressions. *Stat Med*. 2015;34:3133–3143.
- [4] Sullivan KE, Puck JM, Notarangelo LD, et al. USIDNET: a strategy to build a community of clinical immunologists. *J Clin Immunol*. 2014;34:428–435.